

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

5

APPLICATION PAPERS

10

OF

15

ROBERT GRUPE

20

FOR

25

SCANNING COMPUTER FILES FOR SPECIFIED CONTENT

BACKGROUND OF THE INVENTION

Field of the Invention

5 This invention relates to the field of data processing systems. More particularly, this invention relates to data processing systems in which it is desired to scan a plurality of computer files to identify one or more computer files having one or more predetermined characteristics indicative of a computer file having at least some specified content.

Description of the Prior Art

10 It is known to provide anti-virus computer programs and E-mail and data filtering programs. Anti-virus and content filtering computer programs may operate in an on-access mode or an on-demand mode. In an on-access mode, when an access request is made to a computer file, that computer file is subject to scanning for computer viruses, worms, Trojans or other undesired content using a library of data defining such content to be detected. In an on-demand mode, a user initiated or scheduled request may be made to scan a portion of or a complete volume of data with all the computer files contained in that volume being subject to scanning to some degree, even if this is merely to determine they are of a type that does not need scanning.

15 In a similar manner, network transport filtering programs may operate to scan incoming or outgoing E-mail messages and other data streams for undesired content. This may take place as the data is flowing in and out, or alternatively may take place in a scheduled on-demand sense whereby a complete volume of data is scanned as a scheduled or user triggered event.

20 A problem in both the above systems is that the amount of data that needs scanning is rapidly increasing. Furthermore, the number of computer viruses and other undesired forms of content is also rapidly increasing. These factors combine to significantly increase the processing load to carry out such tasks. The situation can arise when the amount of slack time, such as overnight periods or weekend periods, is

insufficient to allow on-demand scans to be run within the period available. This can have the result that such scans need to be terminated early without fully completing thereby potentially significantly compromising the security and usefulness of such systems.

5

SUMMARY OF THE INVENTION

Viewed from one aspect the present invention provides a computer program product comprising a computer program operable to control a master computer to scan a plurality of computer files to identify one or more computer files having one or more
10 predetermined characteristics indicative of a computer file having at least some specified content, said computer program comprising:

scan initiating logic operable in response to a scan initiating command to initiate a scan of said plurality of computer files;

scan dividing logic operable to divide said scan into a plurality of tasks;

15 task issuing logic operable to issue at least some of said plurality of tasks to one or more further computers to be performed by respective ones of said one or more further computers;

task result receiving logic operable to receive from said one or more further computers results of said tasks performed by respective ones of said one or more further
20 computers; and

scan result merging logic responsive to said results of said tasks performed by respective ones of said one or more further computers to identify any of said plurality of computer files having any of said one or more predetermined characteristics to provide a merged scan result.

25

The invention recognises that the concepts of distributed processing may be applied to such scanning for specified content using one or more predetermined characteristic tests. The full scan to be performed may be broken down into separate tasks that can be spread out between more than one computer thereby decreasing the total
30 time required for the scanning. This type of scanning operation is particularly well suited to such distributed processing since it is highly computationally intensive and yet for the

vast majority of time the results produced are null results requiring very little feedback of data to the central source. This makes such a distributed approach quite practical since in the vast majority of cases a block of data that is, for example, passed out for virus scanning will not reveal any viruses thus requiring very little data to be passed back to the co-ordinating source and allow a comparatively simple structure and arrangement of the system.

It will be appreciated that whilst the invention is suitable for scanning for specified content of a variety of types within large numbers of computer files, it is particularly well suited to computer virus/worm/Trojan detection and electronic messaging filtering situations.

The action taken upon detection of a computer file with unwanted content could vary, but typically will include banning that computer file from use, quarantining that computer file or redirecting a message or file for checking elsewhere.

The computer files being scanned may typically reside upon the network drive of a server computer. However, more generally the computer files may be stored on one or more networked data storage devices. It is becoming more common for distributed data storage to be provided within networks. The present techniques could also be applied for scanning the local hard drives of client computers if desired.

The master computer that controls the distributed processing may also perform some of the scanning operations itself. In these circumstances, high priority scanning operations may be kept to be performed by the master computer with lower priority scanning being distributed to further computers. As an example, the master computer may choose to scan for a subset of the total number of computer viruses or worms or Trojans that are known to be the most damaging/common/recent of the total number scanned for and accordingly the ones for which it is most critical to check.

The master computer may preferably be arranged to monitor the performance of the further computers to which the processing is distributed. The master computer may poll the further computers for their reported status in conducting the processing that has been distributed to them and/or may operate a timeout function whereby if a response
5 from a further computer is overdue, then the task that was allocated to that further computer is allocated elsewhere. In such circumstances, diagnostic operation may be commenced with the further computer to determine whether further tasks should be sent to that computer.

10 With the advent of "PC on a card" type computers, it is possible for the one or more further computers that co-operate with the master computer to be provided connected to the system bus of the master computer thereby achieving high efficiency distributed processing.

15 Viewed from another aspect the invention provides a computer program product comprising a computer program operable to control a computer to scan a plurality of computer files to identify one or more computer files having one or more predetermined characteristics indicative of a computer file having at least some specified content, said computer program comprising:

20 task receiving logic operable to receive a task request from a master computer specifying a portion of a larger scan operation;
task performing logic operable to perform said portion of a larger scan operation;
and
result reporting logic operable to report a result of said portion of a larger scan
25 operation to said master computer.

It will be appreciated that the computer program executing on the master computer to initiate and co-ordinate the distributed scanning together with the computer program executing on the further computers together represent complementary aspects of
30 the same invention.

In order to improve the reliability and security of the scanning, the further computers may be arranged to provide update status information to the master computer (possibly upon request by the master computer) such that the master computer can determine if the further computer has the latest data for identifying the specified content required for the scan.

Further computers may be made available in accordance with a predetermined schedule or possibly on a more ad-hoc basis. Scheduled availability could be arranged via scheduling on the further computer or by request from the master computer. An ad-hoc approach could be catered for by an availability notification being issued from the further computers to the master computers, such as upon commencement of execution of a screen saver program indicating that the computer was not in active use by a user.

Complementary aspects of the invention also provide a method of controlling a computer and a computer apparatus.

The above, and other objects, features and advantages of this invention will be apparent from the following detailed description of illustrative embodiments which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a computer network;

Figure 2 illustrates is a flow diagram illustrating the processing of a master computer controlling distributed scanning;

Figure 3 is a flow diagram illustrating processing within a further computer taking part in distributed scanning; and

Figure 4 schematically illustrates a general purpose computer of a type that may be used to perform scanning operations.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Figure 1 illustrates a computer network 2 including a file server 4, a mail server 6, a plurality of client computers 8, 10, 12 and a plurality of rack mounted appliance computers 14. A local area network 16 connects these computers.

The file server 4 includes a network storage device 18 providing file storage for the computers connected to the local area network 16. The file server 4 also includes two computers 20 implemented as PC computers on cards connected to the system bus of the file server 4.

In operation, the network storage device 18 is subject to regular on-demand scans to identify computer viruses, Trojans or worms. As the network storage device 18 can be very large, the amount of processing to compare every file stored against a large and ever growing number of virus definition profiles can also be very large. This can result in insufficient time being available to complete a full on-demand scan during the scheduled inactive periods of the network. As another example, the mail server 6 may include a network storage device storing all the received and sent E-mails into and out of the network and these may be subject to on-demand content filtering during quiet periods of network activity. A further example may be that the local hard disk drives of the client computer 8, 10, 12 may be subject to virus scanning co-ordinated and driven by the file server 4 during quiet network periods providing that those local hard disk drives are shared and available for access over the network.

Figure 2 is a flow diagram illustrating the processing performed by the file server 4 in accordance with one example of the invention (it will be understood that any master control computer on the network could perform this role). At step 22 a file server starts an on-demand scan on the network storage device 18. This may be triggered as a scheduled event or in response to a specific user input. The on-demand scan could be triggered in other ways, for example a combination of factors such as low network activity and in excess of a minimum period since the last on-demand scan.

At step 24, the file server 4 identifies those computers on the network 16 that are available to assist with the scanning operations. The availability of a computer for taking part in the distributed processing may be detected by issuing a polling request to all connected computers. The connected computers could themselves be operating in accordance with a schedule that indicates that they were available for such use during predetermined known quiet times. Alternatively, a screen-saver operating or other idling mode indicating parameter (such as activation of a power saving mode) could be used as indicating that a particular computer was available to take part in the scanning activity. Step 24 also serves to check that the computers indicated as available for the scanning operation have the latest copy of the virus definition data or content filtering data available to them. If a particular computer indicates an out-of-date set of data, then updated data may be downloaded from the file server 4 or a remote site, such as a virus program provider's FTP site.

At step 26, the file server (master computer) 4 checks to see if any computers 8, 10, 12, 14, 20 that are available for taking part in the distributed processing are available to have a task allocated to them. If any such computers are available, then step 26 issues a scanning task to them. This scanning task may be to allocate a predetermined block of computer files on the network storage device 18 to that computer to be scanned for either computer viruses, unwanted computer programs, or restricted information contents depending upon the circumstances. The file server 4 may be arranged to itself check all of the computer files on the network storage device 18 for the highest priority items such as the latest released viruses, the most damaging viruses or other high priority unwanted content. In these circumstances, the bulk of the lower priority scanning may be offloaded to the distributed further computers on the network 2. At step 30, the file server 4 checks to see if any of the further computers to which tasks have been issued is waiting to send results back to the file server 4. If results are ready to be sent back, then these are collected at step 32 and any necessary action in response to detected problem files is taken at step 34.

If no results are awaiting collection, then processing proceeds to step 36 at which all of the computers who have been issued tasks are polled to respond to indicate that they are

all still processing their allocated tasks normally, even though they have not yet finished. If any of the further computers indicates a status showing that normal processing is not continuing, then control passes to step 38 for remedial action.

5 If all the poll results indicate normal processing, then step 40 checks to see if any of the computers is overdue for returning its results. It may be that a computer will respond normally to a polling request, but nevertheless be malfunctioning, and in this case the overdue nature of the result may be used to trigger remedial processing by step 38.

10 Step 38 is triggered when a task is not completing normally by a further computer and serves to return the allocated task to the list of tasks still to be performed. At step 42, the further computer may be subject to diagnostic exchanges to test whether it is fit to have further tasks allocated to it, and if not, the further computer may be removed from the list of further computers available to take part in the distributed processing.

15 At step 44, the issuing of tasks, the collection of results, and the testing of on-going tasks has all been performed by the file server 4 and accordingly the file server 4 can devote some processing resources to carrying out some of the scanning activities itself. These may be the highest priority and most critical scanning operations as previously mentioned. If any
20 problem files are identified by this local scanning, then step 46 serves to trigger appropriate problem file action at step 48. The problem file action taken at step 48 may be the same as the problem file action taken at step 50 following a problem file detected by a further computer at step 32.

25 Step 52 serves to determine whether or not the scanning is complete. If further scanning operations are still required, then processing returns to step 26. Otherwise, the processing terminates.

30 Figure 3 illustrates the processing that may take part in one of the further computers 8, 10, 12, 14, 20 illustrated in Figure 1. At step 54, the further computer registers with the file server 4 as being available to take part in the scanning operations. This registration may

be triggered as a result of a scheduled operation or alternatively may be triggered through starting of a screen saver program or through some other detected parameter. The further computer waits at step 56 until a scanning task is received from the file server 4. When a scanning task is received, then the further computer reads the files it is to scan from the network storage device 18 at step 58 and stores these locally. At step 60 it runs the scanning operations to detect the specified content (such as the presence of virus programs, worm programs, Trojan programs or the like or undesired content). When this scanning is completed, then processing waits at step 62 for the results to be collected from the further computer by the file server 4. At step 64, the further computer tests as to whether or not it should terminate its participation in the distributed processing, such as its scheduled availability coming to an end or its screen saver operations ceasing. If the further computer is to continue to participate, then processing returns to step 56. Otherwise, processing terminates.

Figure 4 illustrates a general purpose computer 200 of the type that may be used to perform the above described techniques. The general purpose computer 200 includes a central processing unit 202, a read only memory 204, a random access memory 206, a hard disk drive 208, a display driver 210 with attached display 211, a user input/output circuit 212 with attached keyboard 213 and mouse 215, a network card 214 connected to a network connection and a PC computer on a card 218 all connected to a common system bus 216. In operation, the central processing unit 202 executes a computer program that may be stored within the read only memory 204, the random access memory 206, the hard disk drive 208 or downloaded over the network card 214. Results of this processing may be displayed on the display 211 via the display driver 210. User inputs for triggering and controlling the processing are received via the user input/output circuit 212 from the keyboard 213 and mouse 215. The central processing unit 202 may use the random access 206 as its working memory. A computer program may be loaded into the computer 200 via a recording medium such as a floppy disk drive or compact disk. Alternatively, the computer program may be loaded in via the network card 214 from a remote storage drive. The PC on a card 218 may comprise its own essentially independent computer with its own working memory, CPU and other control circuitry that can co-operate with the other elements in Figure 4 via

the system bus 216. The system bus 216 is a comparatively high bandwidth connection allowing rapid and efficient communication.

It will be appreciated that the above computer apparatus is only one example of the computer platforms that may operate in accordance with the present techniques. As examples, handheld computers and interactive televisions may be used. The display driver may also serve to provide audio output and the input devices may include pen based devices, cameras (for automated recognition) and microphones. The hard disk drive may be replaced or supplemented with a floppies disk drive, a CD/DVD drive or memory cards. It will be understood that these are only some examples of the variation in computer platform that may be used as appropriate.

Although illustrative embodiments of the invention have been described in detail herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various changes and modifications can be effected therein by one skilled in the art without departing from the scope and spirit of the invention as defined by the appended claims.